HANDWRITTEN CHARACTER RECOGNITION USING*NAÏVE BAYESCLASSIFIER* METHOD FOR DESKTOP AND MOBILE APPLICATION

S. Priyanthy and HMM. Naleer

Department of Mathematics, Faculty of Applied Sciences, South Eastern University of Sri Lanka
selvarajahpri@gmail.com, hmmnaleer@seu.ac.lk

Abstract

In Recent years, character recognition has gained more importance in the area of pattern recognition owning to its application in various domains. Many OCRs systems are been applied, but less interest have been given to document images obtained by camera phone.Off-line recognition of handwritten words is a difficult task due to the high variability and uncertainty of human writing. This paper presents a complete offline handwritten recognition system which describes the implementation of a desktop application and an android application. Our system includes five stages namely: pre-processing, segmentation, feature extraction, classification and postprocessing. Input for this system is a photo of handwritten text captured by a camera phone. Then it was directed to above stages and finally the output is produced. Naïve Bayes (NB) classification algorithm is used as classifier. In classification process we cut the image in several blocks. For each block, we compute a vector of descriptors. Then, we use K-means to cluster the low-level features including Zernike and Hu moments. Finally, we apply Bayesian networks classifier to classify the whole image of words. Experiments were performed with handwritten and machine-printed character images. The results indicate that the proposed system is very effective and yields good recognition rate for character images obtained by camera.

Keywords:OCR, handwriting recognition, Naïve Bayes (NB) classifier

Introduction

Optical Character Recognition (OCR) is an interesting and challenging field of research in pattern recognition, artificial intelligence and machine vision and is used in many real life applications processing, vehicle number plate recognition, tax forms processing, and digit recognition. A lot of research work has been done in this field considering the scope of the area [1]. OCR represents the mechanical or electronic translation of images containing handwritten, typewritten or printed text (usually captured by a scanner) into computer editable text. Such technique is to facilitate smoother interaction between man and machine. The technique of handwritten recognition can contribute tremendously to the development of a complete OCR system. Therefore OCR of handwritten numerals is still an active area of research [7]

Handwriting recognition (HWR) is a mechanism for transforming the written text into a symbolic representation, which plays an essential role in many human–computer interaction applications [2].

In general, HWR can be categorized into two distinct types: online and off-line based systems. Online recognition is relatively easier as it can make use of additional information not available to the off-line systems such as the strength and sequential order of the writing. On the contrary, off-line recognition is more difficult as it is based solely on images of written texts. However, online recognition is impossible in many applications hence off-line recognition is focused in this paper.

On-line HWR involves the automatic conversion of text, as it is written, on a special digital device or PDA, where a sensor captures the movements of the writing instrument and also the pen-up/pen-down switching. This kind of data is known as digital ink and pcan be regarded as a dynamic representation of handwriting. The obtained signal is converted into letter codes which can be used by the computer and in text processing applications. On-line character recognition is sometimes confused with optical character recognition.

Off-line HWR involves the automatic conversion of text from an image into letter codes that can be used within a computer or other text processing applications. The data obtained in this form is regarded as a static representation of handwriting. This technology is successfully used in businesses that process a large amount of handwritten documents, such as insurance companies. The off-line HWR is difficult, because different people have different handwriting styles. Nevertheless, limiting the types of input data allows improving the recognition process [8]

The recognition of handwritten words can be divided into segmentation based or segmentation free approaches. The former segments words into characters or letters for recognition and can be regarded as an analytical approach. The latter, which can be regarded as a global approach, takes the whole word image for recognition and therefore needs no segmentation. Although the global approach makes the recognition process simpler, it requires a larger input vocabulary than analytical approach, hence segmentation based approach is focused in this paper. [3]

Feature extraction is to remove the redundancy from the data and gain a more effective representation of the word image by a set of numerical characteristics, i.e. Extracting most essential information from raw images. According to features used in off-line recognition are classified into high level features which are extracted from the whole word image, medium level features which are extracted from the letters, and low level features which are extracted from sub-letters, The clustering of low level features of the image is given [3].

Several OCR systems have been proposed by researchers, but fewer attentions have been given to document image recognition acquired via camera phone. Therefore our objective is mainly interested in the development of an off-line English handwriting and machine-printed characters recognition system, in which the images are obtained by camera phone. The ultimate goal of this project is to build a recognizer that can automatically recognize (correctly) an unseen word with a relatively high certainty.

Input image is obtained by camera or taking a machine-printed character images and allowed to go through several stages. Those stages are namely: pre-processing, segmentation, feature extraction, classification and postprocessing.In the process of classification we cut the image in several blocks. For each block, we compute a vector of descriptors. Then, we use K-means to cluster the low-level features including Zernik and Hu moments. Finally, we apply Bayesian networks classifier to classify the whole image of words. Multilayer Perceptron also used in this implementation which is a special kind of Artificial Neural Network (ANN). ANN is developed to replicate learning and a generalization ability of human’s behaviour with an attempt to model the functions of biological neural networks of the human brain. Nowadays MLP is the mostly used classifier in the field of handwritten character recognition. Both desktop application and an android application have been developed.The remainder of the paper is organized as follows: chapter 2 is literature review and chapter 3 explains about methodology, then chapter 4 explains the system overview, chapter 5 explains the results and discussion part and finally chapter 6 explains conclusion of this paper.

Jawad H AlKhateeb‘s research is multi-class classification system is of handwritten Arabic words using Dynamic Bayesian Network (DBN) is proposed, in which technical details are presented in terms of three stages, i.e. pre-processing, feature extraction and classification. This paper proposed a system to classify Arabic handwritten word using DBN based on density features using sliding window technique. The system has been applied to the IFN/ENIT database of handwriting words written by different writers. Employing DBN as a recognition tool in handwritten Arabic word recognition system has shown good results which produces good rate of recognition rates. The proposed system relies on the DBN classification and density features. This system can be applied to other patterns with slightly adaptation.There are three main drawbacks for the DBNs used in this research:
1) Long training time required for some applications is one of the main drawbacks of DBN.
2) The DBN is not suitable for an imbalanced data such as the IFN/ENIT database. To overcome this drawback, a random reading for the images in each folder is done.
3) The entire images have to be equal in size.[2][3]

Elie Krevat and Elliot Cuzzillo used Naïve Bayes Algorithm as their classifier for the research “Improving Off-line Handwritten Character Recognition with Hidden Markov Models” Their results; first and foremost, demonstrate that a Naive Bayes classifier can perform remarkably well on images of handwritten characters even though the Naive Bayes assumption is not wholly accurate. The Naive Bayes classifier accounted for a significant portion of their solution’s accuracy, without applying any complex feature extraction or pre-processing on the data set. Their results also show a significant benefit derived from exploiting correlations between adjacent letters in words when a dictionary is infeasible to create or unavailable for the testing set.

Methodology

HWR is a challenging task because of variability of writing styles of different writers from different environment. Offline HWR is significantly different from online HWR, because here, stroke information is not available [6].
The task becomes more tedious when the text document quality is low and if the characters are written very close to each other. Some characters have similar shapes that require advanced and complex techniques for recognition. Any character recognition system has number of number of descriptive stages. Below the general steps of character recognition process are explained.

Data acquisition

The input to the OCR system is the scanned document image. This input image should have specific format such as .jpg. This image is acquired through a scanner, digital camera or any other suitable digital input device. After image acquisition the image data goes through following processes.

Pre-processing

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately.The main objectives of pre-processing are:

- Binarization- Document image binarization (thresholding) refers to the conversion of a grey-scale image into a binary image.
- Noise reduction (morphological operators) - removes isolated specks and holes in the characters. Noise reduction improves the quality of the document.
- Normalization- This stage removes some of the variations in the image that do not affect the identity of the input data and provides a tremendous reduction in data size.

Segmentation

Segmentation is the most important aspect of the pre-processing stage. It allows the recognizer to extract features from each individual character. In the more complicated case of handwritten text, the segmentation problem becomes much more difficult as letters tend to be connected to each other, overlapped or distorted.

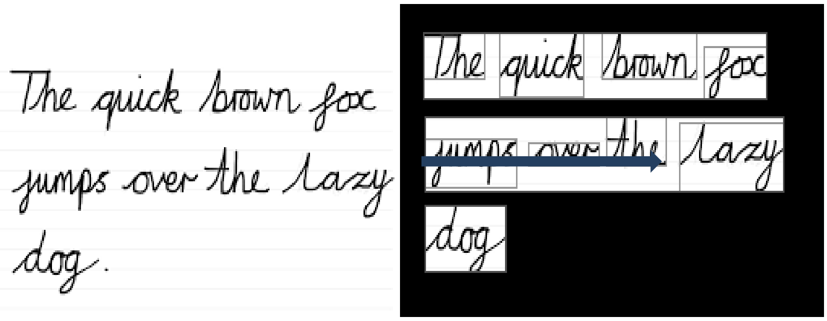


Figure 1. Example of Segmentation

Feature extraction

In feature extraction stage, each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements and to generate similar feature set for variety of instances of the same symbol.

Classifications

Feature extraction stage gives us the feature vector that is used for classification. Classification is the decision making step in the OCR system that makes use of the features extracted from the previous stage in the process.

To do the classification we must have a data bank to compare with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of data bank. We have used naïve Bayes classifier and multi-layer perceptron classifier in our project.

Post-processing

The purpose of this step is the incorporation of context and shape information in all the stages of OCR systems. It is necessary for meaningful improvements in recognition rates. A dictionary can be used to correct minor errors.

System Overview

The proposed system comprises three main modules: a module for extracting primitive blocks from the local cutting of words images, a classification module of primitives in the cluster by using the method of k-means, and a classification module the overall picture based on the structures of Bayesian networks developed for each class. The following sub-sections describe the different step of the proposed system.

Step 1: decomposition of images into blocks

Classifying an image depends on classifying the objects within the image. The studied objects in an image are the following analysis represented by blocks.

Step 2: Features Extraction

After the pre-treatment of the word, we use the descriptors such as moment invariants of Zernike and Hu descriptors, which are invariant to rotation, translation and scaling, to extract the characteristics of each block.

Step 3: Clustering of blocks with K-means

Approach is based on modelling of the image by three blocks reflecting a local description of the image. At each block we will first apply the descriptors presented in the previous section and second we classify it by Bayesian classifier. To generate the label vector from vector descriptor used the k-means algorithm. Each vector will undergo a clustering attribute, and replace the labels generated vector components descriptors. We use the method of k-means to cluster the descriptor.

Step 4: Structure Learning

The originality of this work is the development of a Bayesian network for each block. Then, each word image is characterized by three Bayesian networks. We propose three variants of Bayesian Networks such as Naïve Bayesian Network (NB), Tree Augmented Naïve Bayes (TAN) and Forest Augmented Naïve Bayes (FAN). [4][5]

Step 6: Classification

In this work the decisions are inferred using Bayesian Networks. Class of an example is decided by calculating posterior probabilities of classes using Bayes rule. This is described for both classifiers. Suppose that C^i class is composed with three subclasses C_t^i linked in over time (t=1, 2, 3). One image is divided on three blocks. Start to classify the blocks B_k with their attributes using Bayes rules:

- The blocks B_{k1} at time t=1 considering C_1^i classes:

$$P\ (C_1^i\ |B_{k1}) = p(C_1^i\ |\ A_{1b1}, \ldots, A_{mb1}) = \frac{p\left(A1b1, \ldots, Amb1|C_1^i\right) * p\ (C_1^i)}{p(A1b1, \ldots, Amb1)} = \frac{\prod_j p\left(Ajb1|C_1^i\right) * p\ (C_1^i)}{p(A1b1, \ldots, Amb1)}$$

- The blocks B_{k2} at time t=2 considering C_2^i classes:

$$P\ (C_2^i\ |B_{k2}) = p(C_2^i\ |\ A_{1b2}, \ldots, A_{mb2}) = \frac{p\left(A1b2, \ldots, Amb2|C_2^i\right) * p\ (C_2^i)}{p(A1b2, \ldots, Amb2)} = \frac{\prod_j p\left(Ajb2|C_2^i\right) * p\ (C_2^i)}{p(A1b2, \ldots, Amb2)}$$

- The blocks B_{k3} at time t=3 considering C_1^i classes:

$$P\ (C_3^i\ |B_{k3}) = p(C_3^i\ |\ A_{1b3}, \ldots, A_{mb3}) = \frac{p\left(A1b3, \ldots, Amb3|C_3^i\right) * p\ (C_3^i)}{p(A1b3, \ldots, Amb3)} = \frac{\prod_j p\left(Ajb3|C_3^i\right) * p\ (C_3^i)}{p(A1b3, \ldots, Amb3)}$$

Note that we do not need to explicitly calculate the denominators:

$$p(A1b1\,,\dots,Amb1),\ p(A1b2\,,\dots,Amb2),\ p(A1b3\,,\dots,Amb3)$$

They are determinate by the normalization condition. Thus it's sufficient to calculate for each sub class C_t^iits likelihood degree:

$$\prod_j p\big(\text{Ajbt} \mid C_t^i\big) * \ p\,(C_t^i)$$

We move to classify the whole image from its entire three blocks considering C^{i} classes

$$\text{P}\ (\text{C}^{\text{i}} \mid \text{image}_{\text{k}}) = \text{p}(\text{C}^{\text{i}} \mid \text{B}_{\text{k1}}\ ,\text{B}_{\text{k2}},\ \text{B}_{\text{k3}}) = \tfrac{1}{3}\sum_{t=3}^{3} p(C_t^i \mid \text{Bkt})$$

To determine the subclass of each block the Bayesian approach do the comparison between the conditional probability $p(C_t^i \mid \text{Bkt})$for each time t and chose the maximum value by applying the Bayes decision rule as follows :

$$\text{d(x)} = \text{argmax}\ C_t^i p(C_t^i \mid \text{Bkt}) = \text{argmax}\ C_t^i \prod_j \text{p}\big(\text{ Ajbt} \mid C_t^i\big) * \ \text{p}\ (C_t^i)$$

To find the image class C^{i} considering classes the Bayesian method do the comparison between the conditional probabilities $\text{p}(\text{C}^{\text{i}} \mid \text{image}_{\text{j}})$.

$$\text{d(x)} = \text{argmax}\ c^{\text{i}}\, \text{p}(\text{C}^{\text{i}} \mid \text{image}_{\text{k}}) = \text{argmax}\ c^{\text{i}}\tfrac{1}{3}\sum_{t=1}^{3} p(C_t^i \mid \text{Bkt})$$

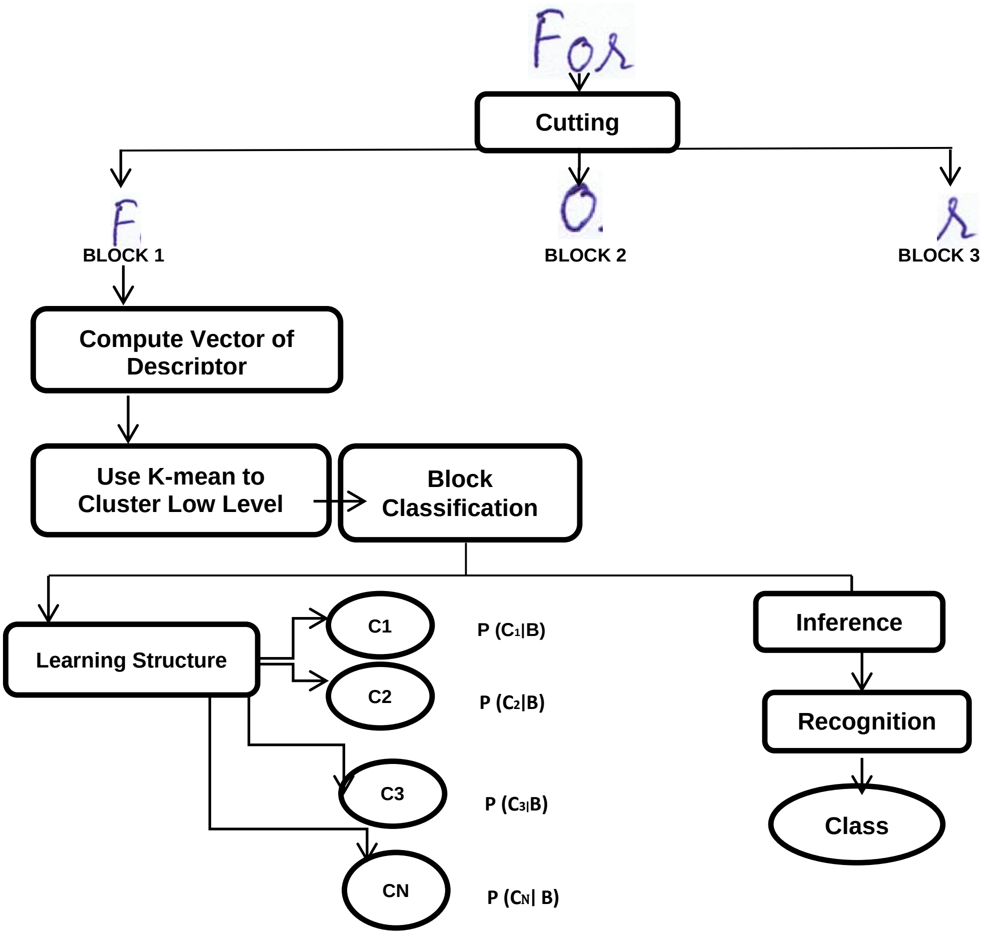


Figure 2: System Overview

Naïve Bayes

A variant of Bayesian Network is called Naïve Bayes. Naïve Bayes is one of the most effective and efficient classification algorithms**.** The conditional independence assumption in naïve Bayes is rarely true in reality. Indeed, naive Bayes has been found to work poorly for regression problems, and produces poor probability estimates. One way to alleviate the conditional independence assumption is to extend the structure of naive Bayes to represent explicitly attribute dependencies by adding arcs between attributes.

It is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real world situations.

Figure 3 shows graphically the structure of naïve Bayes, each attribute node has the class node as it parent, but does not have any parent from attribute node. As the values of P (a_i |c) can be easily calculated from training instances, naïve Bayes is easy to construct.

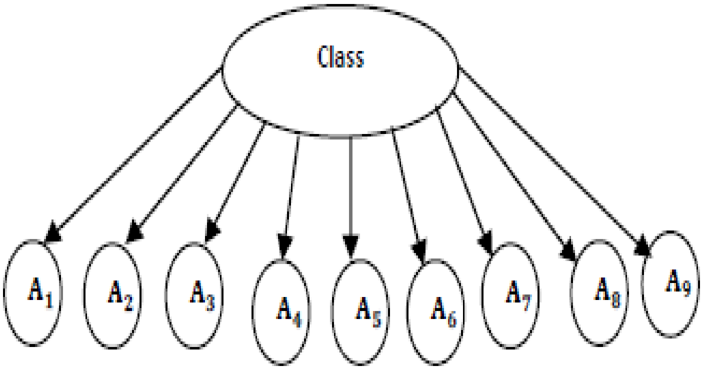
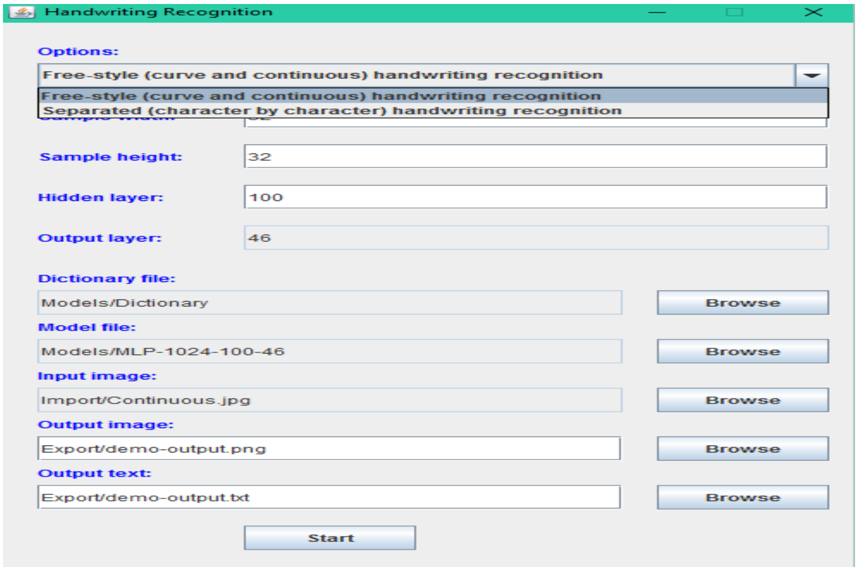


Figure 3: Naïve Bayes (NB)

Results and Discussion

When run the project, we have to select whether we want to recognize separated handwriting or continuous handwriting using drop down menu in the main frame which shown below. And also we have the option to select input image (photo of handwritten text taken by camera phone) which has to be recognized. After selecting all necessary inputs we can press start button which brings us to the output of recognized word will be display in a window called output.txt.

In android application for mobile phones the output produced in the form of .apk extension. We can install this application in any android mobile phones using normal installation procedure. In this offline OCR system, firstly we have to train the digit from 0-9, then we can draw a digit and if we press the button Recognize then the output will be displayed. We used Naïve Bayes as classifier, therefore we can notice that output displayed with probability.



Options:

Free-style (curve and continuous) handwriting recognition

Free-style (curve and continuous) handwriting recognition
Separated (character by character) handwriting recognition

Sample height: 32

Hidden layer: 100

Output layer: 46

Dictionary file: Models/Dictionary

Model file: Models/MLP-1024-100-46

Input image: Import/Continuous.jpg

Output image: Export/demo-output.png

Output text: Export/demo-output.txt

Figure 4: Main Frame

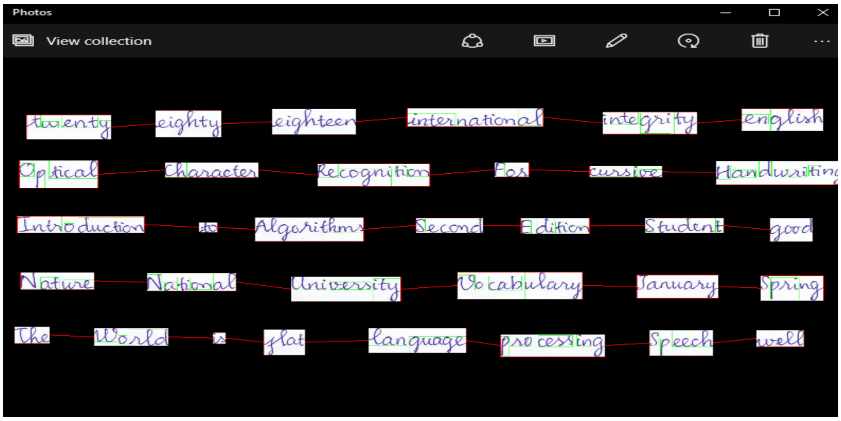
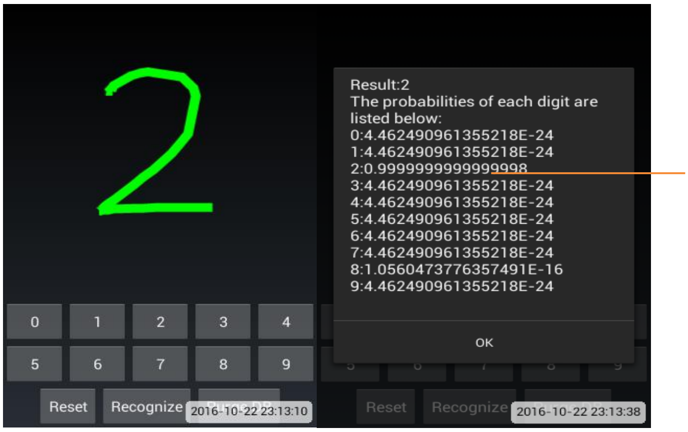


Figure 5: segmented words

There are several problems in the use of Bayesian networks. The first one is the correspondence between the graphical structure and associated probabilistic structure will allow reducing all the problems of inference problems in graph theory. However, these problems are relatively complex and give rise to much research.

The second difficulty of Bayesian networks lies precisely in the operation for transposition of the causal graph to a probabilistic representation. Even if the only probability tables needed to finish the entire probability distribution are those of a node conditioning compared to his parents, he is the definition of these tables is not always easy for an expert.

Another problem of Bayesian networks, the problem of automatic learning of the structure that remains is a rather complex problem.



```
Result:2
The probabilities of each digit are
listed below:
0:4.462490961355218E-24
1:4.462490961355218E-24
2:0.9999999999999998
3:4.462490961355218E-24
4:4.462490961355218E-24
5:4.462490961355218E-24
6:4.462490961355218E-24
7:4.462490961355218E-24
8:1.0560473776357491E-16
9:4.462490961355218E-24
```

(a) (b)

Figure (a & b) 6: testing the trained classifier

Conclusion

This objective of this study was to try to build a word recognizer that classify a handwritten text which captured by a camera phone. Wehave successfully implemented the desktop and android application using Naïve Bayes algorithm as our classifier.Naive Bayes being an easy, fast and most of the times "quite accurate", it is also "Naive" because it makes the assumption of conditional independence of the features.

The Bayesian Classifier is capable of calculating the most probable output depending on the input. It is possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. And weused Multi Layer Perceptron also in our project.

An artificial neural network consists of a number of processors very simple and strongly interconnected, called neurons, which are analogue to the biologic neurons of human brain. The main property of artificial neural networks is the capacity of learning. The multilayer perceptron (MLP) is the most widely known and used type of neural network By training it means to train them on particular inputs so that later on we may test them for unknown inputs (which they have never seen before) for which they may classify or predict etc. (in case of supervised learning) based on their learning.

During our project we faced a lot of difficulties, it is mainly due to the fact that numerous variations in writing styles of individuals. And also image quality also affects the recognition rate. However using Naïve Bayes classifier the resultswe got is quite accurate.

References

[1] Gaurav Y. Tawde, (February 2014), "Optical Character Recognition for Isolated Offline Handwritten Devanagari Numerals Using Wavelets", Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 2(Version 1), pp.605-611

[2] Jawad H. AlKhateeb, Olivier Pauplin, Jinchang Ren, Jianmin Jiang, “Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition”, Knowledge-Based Systems 24 (2011) 680–688.

[3] Jawad H AlKhateeb, Jinchang Ren, Jianmin Jiang, Husni Al-Muhtaseb , (2011) , “Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking”,Pattern Recognition Letters 32 1081–1088.

[4] Jayech K , Mahjoub M.A. (2010) New approach using Bayesian Network to improve content based image classification systems, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6.

[5] Jayech K , Mahjoub M.A (2011) . Clustering and Bayesian Network to improve content based image classification systems, International Journal of Advanced Computer Science and Applications- a Special Issue on Image Processing and Analysis.

[6] Munish Kumar, M. K. Jindal, R.K. Sharma,(November, 2011),“k-nearest neighbor based offline handwritten Gurumukhi character recognition,” International Conference on Image Information Processing 2011(ICIIP), Published by IEEE Computer Society, Jaypee University of Information Technology, Waknaghat, Shimla, Himachal Pradesh, India, (ISBN No. 978-1-61284-860-0),pp. 1-4, 3-5.

[7] Nibaran Das, (2006) , Ayatullah Faruk Mollah, Sudip Saha, Syed Sahidul Haque, “Handwritten Arabic Numeral Recognition using a Multi-Layer Perceptron”, Proc. National Conference on Recent Trends in Information Systems 200-203.

[8] Violeta Sandu And Florin Leon, (2009),” Recognition Of Handwritten Digits Using Multilayer Perceptrons”, BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI Publicat de Universitatea Tehnică „Gheorghe Asachi” din Iaşi Tomul LV (LIX), Fasc. 4, Secţia AUTOMATICĂ şi CALCULATOARE